[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G11B 7/006
G11B 7/007 G11B 20/10

[12] 发明专利申请公开说明书

[21] 申请号 02148818.5

[43] 公开日 2004 年 6 月 2 日

[11] 公开号 CN 1501364A

[22] 申请日 2002.11.18 [21] 申请号 02148818.5

[71] 申请人 华为技术有限公司
地址 518057 广东省深圳市南山区科技园科发路 1 号华为用服中心大厦
[72] 发明人 张 巍 张国彬 任雷鸣 陈绍元 郑 珉 胡 鹏

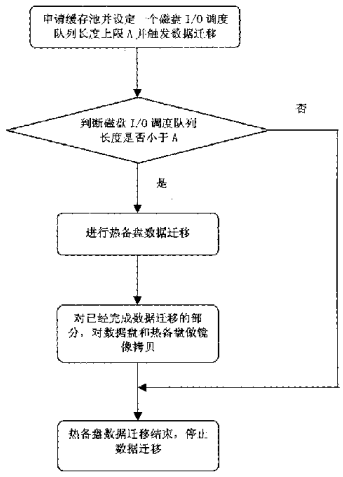
[74] 专利代理机构 北京三友知识产权代理有限公司
代理人 李 强

权利要求书 2 页 说明书 7 页 附图 3 页

[54] 发明名称 一种热备盘数据迁移方法

[57] 摘要

一种热备盘数据迁移方法，涉及磁盘存储领域。包括以下步骤：a、设定一个磁盘 I/O 调度队列长度上限 A 并触发数据迁移；b、判断磁盘 I/O 调度队列长度是否小于 A，如果是，进行热备盘数据迁移，如果否，则停止数据迁移；c、对已经完成数据迁移的部分，对数据盘和热备盘做镜像拷贝，直至热备盘数据迁移结束。本发明在保持热备盘位置固定的同时，实现 RAID 组磁盘的统一管理。



ISSN1008-4274

1、 一种热备盘数据迁移方法，其特征在于包括以下步骤：

a、申请缓存池并设定一个磁盘 I/O 调度队列长度上限 A，触发数据迁移；

b、判断磁盘 I/O 调度队列长度是否小于 A，如果是，进行热备盘数据迁移，如果否，则停止数据迁移；

c、对已经完成数据迁移的部分，对数据盘和热备盘做镜像拷贝，直至热备盘数据迁移结束。

2、 如权利要求 1 所述的热备盘数据迁移方法，其特征在于在所述的缓存池包括自由缓存和预留缓存。

3、 如权利要求 1 或 2 所述的热备盘数据迁移方法，其特征在于在上述数据迁移的过程中，热备盘数据迁移 I/O 调度队列优先级小于主机读写访问 I/O 调度队列优先级。

4、 如权利要求 3 所述的热备盘数据迁移方法，其特征在于在上述数据迁移的过程中，若热备盘正在向某个自由缓存读出数据，如果有主机的读请求，则挂起主机的读请求，直至热备盘该分条单元的数据被完全读到自由缓存中，再恢复被挂起的主机读请求，从自由缓存中直接读取数据，以响应主机对该分条单元的读请求。

5、 如权利要求 4 所述的热备盘数据迁移方法，其特征在于在上述数据迁移的过程中，在响应主机读请求时，热备盘的数据迁移 I/O 调度队列被挂起。

6、 如权利要求 3 所述的热备盘数据迁移方法，其特征在于在上述数据迁移的过程中，若热备盘某个分条单元的数据已完全读到自由缓存中，正在等待数据迁移到数据盘，若此时遇到主机对该分条单元的读请求，则挂起热备盘的数据迁移，从自由缓存中直接读取数据，读取数据之后再恢复热备盘的数

据迁移。

7、 如权利要求 3 所述的热备盘数据迁移方法，其特征在于在上述数据迁移的过程中，如果热备盘的某个分条单元正在进行数据迁移，若此时遇到主机对该分条单元的读请求，则从自由缓存中直接读取数据，以响应主机对该分条单元的读请求。

8、 如权利要求 3 所述的热备盘数据迁移方法，其特征在于在上述数据迁移的过程中，如果热备盘正在向某个自由缓存读出数据，若此时主机对该分条单元写请求，若还有预留的缓存，则直接将需写入的数据写到预留的缓存中；如果预留缓存已耗尽，则先向系统申请一块缓存，将需写入的数据写到该缓存中，热备盘的数据迁移缓存被重定向到新指定的缓存。

9、 如权利要求 3 所述的热备盘数据迁移方法，其特征在于在上述数据迁移的过程中，如果热备盘某个分条单元的数据已完全读到自由缓存中，正在等待数据迁移到数据盘，若此时遇到主机对该分条单元的写请求，则挂起热备盘的数据迁移，将主机所需写入的数据直接写到该自由缓存中，完成后再恢复热备盘的数据迁移过程。

10、 如权利要求 3 所述的热备盘数据迁移方法，其特征在于在上述数据迁移的过程中，如果热备盘的某个分条单元正在进行数据迁移，若此时遇到主机对该分条单元的写请求，若还有预留的缓存，则直接将需写入的数据写到预留的缓存中；若预留缓存已耗尽，则先向系统申请一块缓存，再将需写入的数据写到预留的缓存中，当自由缓存中的数据被迁移完成之后，重新对新指定的缓存中的数据进行数据迁移。

一种热备盘数据迁移方法

技术领域

本发明涉及磁盘存储领域，尤其涉及一种在磁盘阵列系统（RAID：Redundant Array of Inexpensive Disks）中，热备盘数据迁移方法。

技术背景

随着科学技术的飞速发展与计算机技术的普遍应用，人们对存储设备的性能要求越来越高， RAID技术已作为一项成熟的技术广泛的应用于磁盘阵列中。其是通过磁盘阵列与数据条块化方法相结合，以提高数据可用率的一种结构，通过数据镜像实现数据冗余可直接从镜像拷贝中读取数据，系统可以自动地交换到镜像磁盘上， 而不需要重组失效的数据。RAID可分为RAID级别1到RAID级别6， 通常称为: RAID 0， RAID 1， RAID 2， RAID 3，RAID 4， RAID 5。每一个RAID级别都有自己的强项和弱项。"奇偶校验"定义为用户数据的冗余信息，当硬盘失效时，可以重新产生数据。

对一个具有冗余能力的RAID系统，其磁盘阵列包含数据磁盘和热备盘两部分。当磁盘阵列中的某个数据磁盘失效后，RAID系统将启动其重构程序，将失效磁盘中的数据重构到热备盘。由于热备盘通常为所有RAID组成员磁盘（数据磁盘）共享，若经过几次重构之后，没有进行从热备盘到更新的RAID组成员磁盘（数据磁盘）的数据迁移，即当失效数据磁盘被更换后，没有将热备盘上的数据迁移到更换后的数据磁盘中，则对RAID组所有磁盘，包括成员磁盘和热备盘的管理带来极大的不便（热备盘位置不固定，而是散落在成员磁盘中间）。此外，有些数据迁移的方法，比如卡内基·梅隆大学在RAIDFrame采用的方法CopyBack，必须离线进行数据迁移，也就是说，在数据迁移过程中禁止有用户请求下发。这显然与RAID系统要求的24×7小时在线服务不相符的。

目前，由于热备盘数据迁移方法较复杂，还没有见到相应的实现方法。

发明内容

本发明的目的就是提供一种热备盘数据迁移的方法，以解决热备份盘数据迁移的问题。

一种热备盘数据迁移方法，其特征在于包括以下步骤：

a、申请缓存池并设定一个磁盘 I/O 调度队列长度上限 A，触发数据迁移；

b、判断磁盘 I/O 调度队列长度是否小于 A，如果是，进行热备盘数据迁移，如果否则停止数据迁移；

c、对已经完成数据迁移的部分，对数据盘和热备盘做镜像拷贝，直至热备盘数据迁移结束。

所述的缓存池包括自由缓存和预留缓存。

上述数据迁移的过程中，热备盘数据迁移 I/O 调度队列优先级小于主机读写访问 I/O 调度队列优先级。

上述数据迁移的过程中，如果有主机的读请求，则挂起主机的读请求，直至热备盘该分条单元的数据被完全读到自由缓存中，再恢复被挂起的主机读请求，从自由缓存中直接读取数据，以响应主机对该分条单元的读请求。

上述数据迁移的过程中，在响应主机读请求时，热备盘的数据迁移 I/O 调度队列被挂起。

上述数据迁移的过程中，若热备盘某个分条单元的数据已完全读到自由缓存中，正在等待数据迁移到数据盘，若此时遇到主机对该分条单元的读请求，则挂起热备盘的数据迁移，从自由缓存中直接读取数据，读取数据之后再恢复热备盘的数据迁移。

上述数据迁移的过程中，如果热备盘的某个分条单元正在进行数据迁移，若此时遇到主机对该分条单元的读请求，则从自由缓存中直接读取数据，以响应主机对该分条单元的读请求。

上述数据迁移的过程中，如果热备盘正在向某个自由缓存读出数据，若此时

主机对该分条单元的写请求，若还有预留的缓存，则直接将需写入的数据写到预留的缓存中；如果预留缓存已耗尽，则先向系统申请一块缓存，将需写入的数据写到该缓存中，热备盘的数据迁移缓存被重定向到新指定的缓存。

上述数据迁移的过程中，如果热备盘某个分条单元的数据已完全读到自由缓存中，正在等待数据迁移到数据盘，若此时遇到主机对该分条单元的写请求，则挂起热备盘的数据迁移，将主机所需写入的数据直接写到该自由缓存中，完成后再恢复热备盘的数据迁移过程。

上述数据迁移的过程中，如果热备盘的某个分条单元正在进行数据迁移，若此时遇到主机对该分条单元的写请求，若还有预留的缓存，则直接将需写入的数据写到预留的缓存中；若预留缓存已耗尽，则先向系统申请一块缓存，再将需写入的数据写到预留的缓存中，当自由缓存中的数据被迁移完成之后，重新对新指定的缓存中的数据进行数据迁移。

本发明借鉴了面向磁盘重构算法、虚拟非易失Cache的体系结构和设计方案，在保持热备盘位置固定的同时，实现RAID组磁盘的统一管理。

附图说明

图1是本发明数据迁移流程图；

图2是本发明数据迁移过程中，镜像拷贝的示意图；

图3是本发明数据迁移与写请求关系示意图；

图4是本发明数据迁移与读请求关系示意图；

图5是本发明一个数据迁移临界状态主机读请求的处理示意图；

图6是本发明一个数据迁移临界状态主机写请求的处理示意图。

具体实施方式

下面结合说明书附图来说明本发明的具体实施方式。

所谓热备盘数据迁移，即当失效的磁盘被新盘替换后，热备盘将自动把数据迁移到原位置上的数据盘上。该热备盘数据迁移为系统内部实现，除启动时需

要人工触发，迁移过程无需干涉。

采用本发明的热备盘数据迁移方法，如图 1 所示，包括以下步骤:

a、 设定一个磁盘 I/O 调度队列长度上限 A 并触发数据迁移;

该磁盘的数据迁移过程应该在热备盘的空闲时间进行。所谓空闲是指如果在一定时间内（例如50ms）没有磁盘的存取请求，即热备盘的I/O调度队列长度QueueLength为0，磁盘就被认为是处于空闲阶段，数据迁移过程便可以开始。

该过程需要人为地触发，当失效磁盘被更换后，操作员手动触发数据迁移。数据迁移启动后，可人工取消。在整个迁移过程中，时刻记录数据迁移的进度，以便支持断点续传，即当系统断电重启后，可从断点处继续迁移。该整个迁移过程，由CPU处理，并由主控程序控制。

事实上，很难做到热备盘的I/O调度队列长度QueueLength为0，因此，我们设置一个I/O调度队列长度QueueLength的上限A，至于A的取值原则，与热备盘的I/O调度队列最大长度Max_QueueLength有关。一般而言，A可定为Max_QueueLength的10%，或更小，只要满足系统要求都可以。

为尽量使热备盘的数据迁移连续，可将数据迁移安排在访问量不大的时候进行，比如深夜。

热备盘数据迁移时，由于写数据盘的速度必定小于读热备盘速度，所以数据迁移前预先申请一块缓存池来保证读热备盘的持续进行。数据迁移过程可理解为两个独立并行的过程：读热备盘过程和写数据盘过程，读热备盘过程将热备盘的数据按分条单元读取到缓存池中暂存，写数据盘过程将读取结束的缓存数据写入数据盘。

在本发明中，缓存池包括自由缓存和预留缓存两部分，自由缓存大小为4～5个分条单元大小，用于正常的数据迁移，预留缓存大小为2～3个分条单元大小，用于暂存数据迁移过程中的主机写请求数据。由于在RAID组中，分条单元大小可能不一致，为便于统一内存申请，缓存池中的每块缓存的大小由分条单元最大的分条深度确定。

b、 判断磁盘 I/O 调度队列长度是否小于 A，如果是，进行热备盘数据迁移，

如果否，则停止数据迁移；

当磁盘I/O调度队列长度QueueLength < A时，开始进行热备盘的数据迁移，否则停止数据迁移。

c、 对已经完成数据迁移的部分，对数据盘和热备盘做镜像拷贝，直至热备盘数据迁移结束。

如图 2 所示，随着热备盘数据迁移的进行，对已完成数据迁移（即热备盘的数据已写入数据盘与之相对应的分条单元内）部分的写访问，必须对数据盘和热备盘做镜像拷贝，即将主机发来的写请求，同时发向热备盘和数据盘。这样使得在数据迁移中，数据盘和热备盘始终能保持数据的一致，且热备盘上的数据始终完整保持最新的数据，以便当数据迁移过程中数据盘失效时，热备盘上的数据仍然可用。

如图 3 所示，若主机写请求只发向数据盘，则可用数据将分布于热备盘和数据盘之上，一旦数据盘失效，将导致部分最新数据需要重构才能恢复。

如图 4 所示，对热备盘中尚未做数据迁移的部分以及对已完成数据迁移部分的读访问，则直接对热备盘相应的分条单元进行读写操作。

在上述地数据迁移过程中，可能遇到以下几种临界情形，本发明采取相应的处理方法:

1、主机为读请求:

(a) 当热备盘正在向某个自由缓存读出数据时，若遇到主机对该分条单元的读请求，可暂时挂起主机的读请求，直至热备盘该分条单元的数据被完全读到自由缓存中，再恢复被挂起的主机读请求，从自由缓存中直接读取数据，以响应主机对该分条单元的读请求。在响应主机读请求过程中，热备盘的数据迁移I/O调度队列被挂起。

(b) 热备盘某个分条单元的数据已完全读到自由缓存中，正在等待数据迁移到数据盘，此时，若遇到主机对该分条单元的读请求，可挂起热备盘的数据迁移，从自由缓存中直接读取数据，以响应主机对该分条单元的读请求，然后再恢复热备盘的数据迁移过程。

如图5所示，即是此种情况的示意图，图中

（1）：从热备盘中读出数据到自由缓存中；

（2）：主机的读请求；

（3）：主机的读请求被重定向到自由缓存。

(c) 热备盘的某个分条单元正在进行数据迁移，此时，若遇到主机对该分条单元的读请求，可从自由缓存中直接读取数据，以响应主机对该分条单元的读请求。

2、主机为写请求：

（a） 热备盘正在向某个自由缓存读出数据。此时，若遇到主机对该分条单元的写请求，若还有预留的缓存，则直接将需写入的数据写到预留的缓存中；若预留缓存已耗尽，则先向系统申请一块缓存，再将需写入的数据写到该缓存中。而该分条单元所用的自由缓存中的数据作废。热备盘的数据迁移缓存被重定向到新指定的缓存。

（b） 热备盘某个分条单元的数据已完全读到自由缓存中，正在等待数据迁移到数据盘，此时，若遇到主机对该分条单元的写请求，可挂起热备盘的数据迁移，将主机所需写入的数据直接写到该自由缓存中。然后再恢复热备盘的数据迁移过程。

（c） 热备盘的某个分条单元正在进行数据迁移，此时，若遇到主机对该分条单元的写请求，若还有预留的缓存，则直接将需写入的数据写到预留的缓存中；若预留缓存已耗尽，则先向系统申请一块缓存，再将需写入的数据写到预留的缓存中。当自由缓存中的数据被迁移完成之后，重新对新指定的缓存中的数据进行数据迁移。由于一个RAID组上有M个LUN（Logical Unit Number），而每个LUN上又有N个分条，对应在该RAID组的某个磁盘，则有M × N个分条单元。数据迁移时，就根据该磁盘上的分条单元地址，从小到大依次进行迁移，直至所有分条单元迁移完成。所谓分条单元正在进行数据迁移，是指该分条单元的数据正在被读到缓存，或已被读到缓存，等待被写入相应的数据盘，或正在将缓存中的数据写入相应的数据盘。

如图6所示，即是此种情况的示意图，图中：

（1）从热备盘中读出数据到自由缓存中；

（2）正在进行热备盘数据的迁移；

（3）主机的写请求；

（4）主机的写请求被重定向到预留缓存或重新申请的缓存；

（5）重新进行数据迁移（镜像拷贝）。

在写操作的数据迁移时，必须对热备盘和数据盘对应的分条单元实行镜像拷贝，以确保热备盘和数据盘数据的一致性。

以上所述的热备盘数据迁移缓存数量可设置为 4～5 个，且预留 2~3 个出来备用。由于写数据盘的速度必定小于读热备盘速度，且考虑到外部主机对数据盘的访问，写盘时间将大大高于读盘所用时间。故在系统允许的前提下，可适当增加热备盘数据迁移缓存数量。此外为了应付数据迁移的临界情形，预留 2～3 个缓存专门给处于临界数据迁移分条单元使用。

在热备盘数据迁移时，可先申请一块缓存池，缓存池大小为 4～5 个分条单元大小。由于在 RAID 组中，分条单元大小可能不一致，为便于统一内存申请，缓存池中的每块缓存的大小由分条单元最大的分条深度确定。

本发明借鉴了面向磁盘重构算法、虚拟非易失Cache的体系结构和设计方案；保持了热备盘位置固定，实现RAID组磁盘的统一管理。

以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求书的保护范围为准。

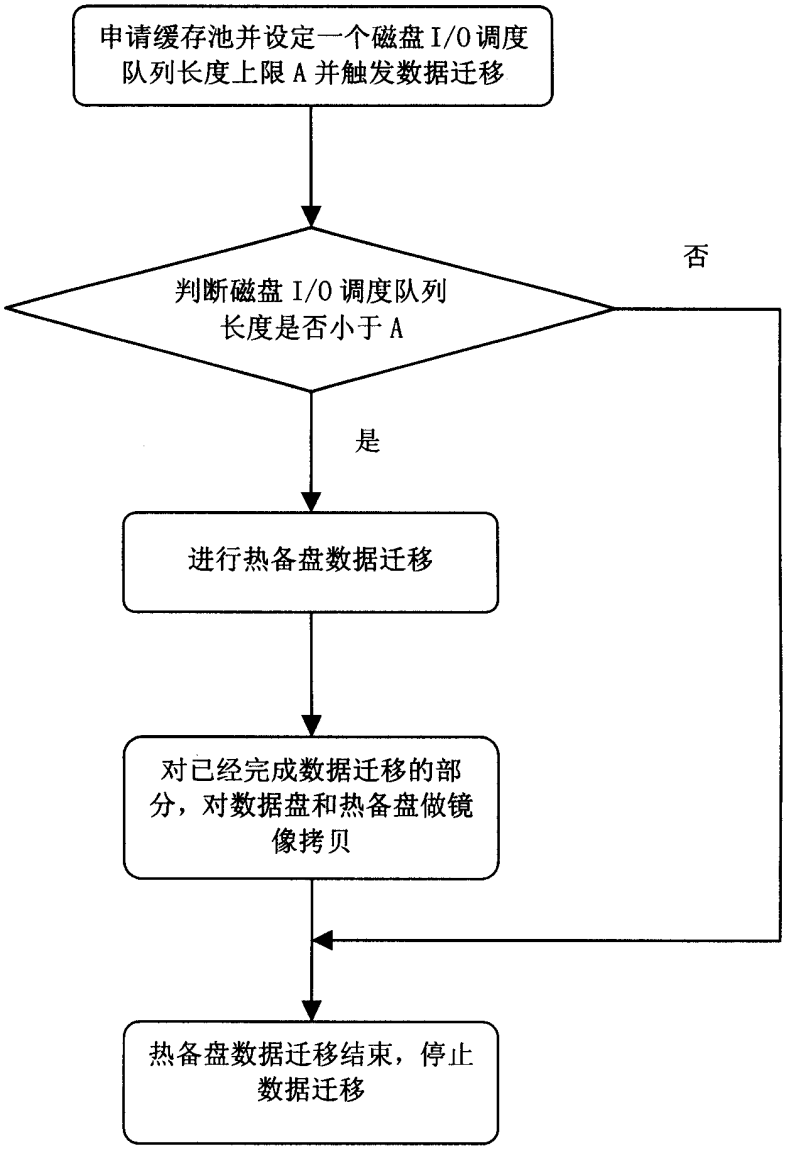

申请缓存池并设定一个磁盘I/O调度队列长度上限A并触发数据迁移
判断磁盘I/O调度队列长度是否小于A
否
是
进行热备盘数据迁移
对已经完成数据迁移的部分，对数据盘和热备盘做镜像拷贝
热备盘数据迁移结束，停止数据迁移

图 1

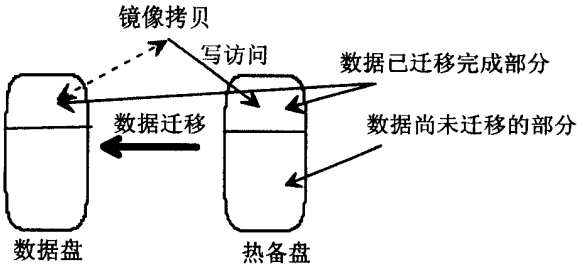

镜像拷贝
写访问
数据已迁移完成部分
数据迁移
数据尚未迁移的部分
数据盘
热备盘

图 2

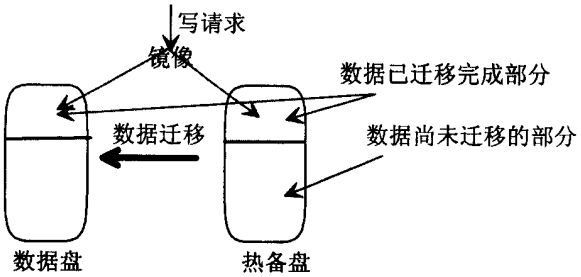

写请求
镜像
数据已迁移完成部分
数据迁移
数据尚未迁移的部分
数据盘
热备盘

图 3

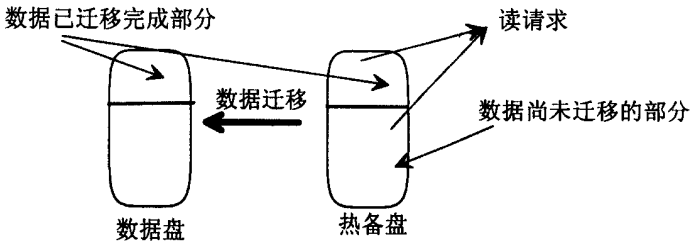

数据已迁移完成部分
读请求
数据迁移
数据尚未迁移的部分
数据盘
热备盘

图 4

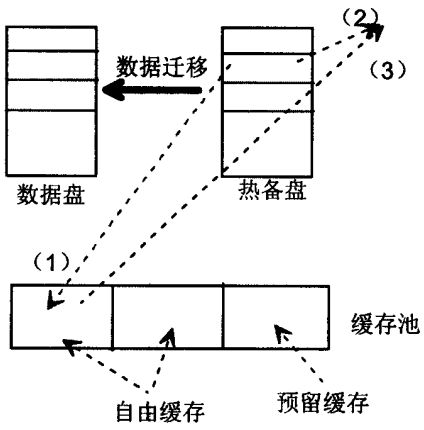

(2)
数据迁移
(3)
数据盘
热备盘
(1)
缓存池
自由缓存
预留缓存

图 5

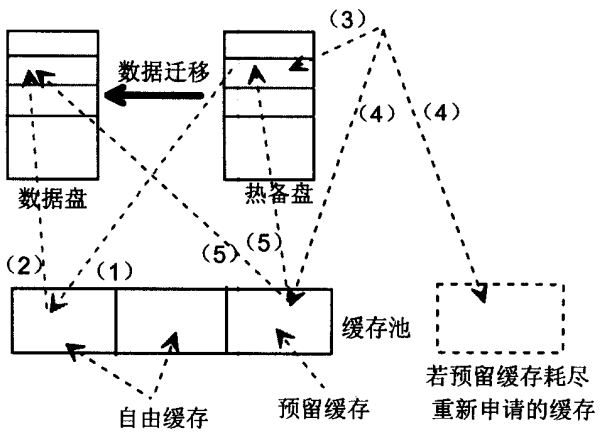

(3)
数据迁移
(4)
(4)
数据盘
热备盘
(5)
(5)
(2)
(1)
缓存池
若预留缓存耗尽
重新申请的缓存
自由缓存
预留缓存

图 6